

S9.2

A 4.8 KBPS MULTI-BAND EXCITATION SPEECH CODER *

John C. Hardwick and Jae S. Lim

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

Recent work has led to the development of a new speech model. This model, referred to as the Multi-Band Excitation (MBE) Speech Model, has been shown to be capable of synthesizing speech without the artifacts common to model-based speech systems. In addition, the MBE speech model has been found to be extremely robust to the presence of background noise in speech. These characteristics make the model particularly useful in the development of high quality speech coding systems. Previous work has demonstrated the advantages of MBE speech coders at bit rates of 9.6 and 8.0 kbps. This paper focuses on the development of a 4.8 kbps speech coding system.

1 Introduction

The problem of analyzing and synthesizing speech has a number of applications, and as a result has received considerable attention in the literature. One class of speech analysis/synthesis systems (vocoders) which have been extensively studied and used in practice are based on an underlying model of speech. For this class of vocoders, speech is analyzed by first segmenting the signal using a window such as a Hamming window. Then, for each segment of speech, the excitation parameters and the system parameters are determined. The excitation parameters consist of a voiced/unvoiced (V/UV) decision and a pitch period. The system parameters consist of the spectral envelope or the impulse response of the system. In order to synthesize speech, the excitation parameters are used to synthesize an excitation signal consisting of a periodic impulse train in voiced regions or a random noise sequence in unvoiced regions. This excitation signal is then filtered using the estimated system parameters.

Even though vocoders based on this underlying speech model have been quite successful in synthesizing intelligible speech, they have not been successful in synthesizing high quality speech. In addition the performance of such a sys-

tem has been observed to degrade rapidly in the presence of background noise. In order to circumvent these limitations a new model was presented in [1]. This Multi-Band Excitation (MBE) speech model replaces the binary voiced/unvoiced classification with a series of voiced/unvoiced decisions. This added degree of freedom allows each speech segment to be partially voiced and partially unvoiced. The result is a speech analysis/synthesis system which is capable of generating high quality speech and which is more robust to the presence of background noise.

One application of the MBE speech model is in speech coding. In [2], an 8.0 kbps MBE vocoder was demonstrated. This system was shown to be capable of high quality reproduction of both clean and noisy speech. The advantage of the MBE speech model was most apparent from the natural quality and the lack of the "buzziness" typically found in vocoded speech. It was postulated that the traditional "buzziness" of vocoder speech is due to replacing noise-like energy in the original speech with periodic energy in the synthetic speech. The MBE vocoder avoids this artifact through added flexibility in the excitation sequence.

In the 8.0 kbps MBE vocoder mentioned above, the model parameters were quantized in a straightforward manner. Experiments showed, however, that substantial redundancy existed amongst the model parameters. Our goal was to choose a coding technique which would utilize these redundancies in order to quantize the parameters more efficiently. In this paper we present the development of a 4.8 kbps MBE vocoder which is designed to accomplish this goal.

In the next section, we review the MBE speech model. In section 3, we briefly discuss the estimation of the model parameters and the synthesis of speech from the estimated model parameters. In section 4, we focus our discussion on the quantization scheme used in developing our 4.8 kbps speech coding system. In section 5, we discuss the performance of our system.

2 Multi-Band Excitation Speech Model

Over a short-time interval, the Fourier transform $S_w(\omega)$ of a windowed speech segment $s_w(n)$ is modeled as the product of a spectral envelope $H_w(\omega)$ and an excitation spectrum $E_w(\omega)$. As in many simple speech models, the spectral envelope is a smoothed version of the original speech spectrum. The excitation spectrum in this new speech model

*This work has been supported in part by the Advanced Research Projects Agency monitored by ONR under Contract No. N00014-81-K-0742, in part by the National Science Foundation under Grant ECS-8407285, and in part by Rome Air Development Center under Contract No. F19628-85-K-0028.

differs from previous models in one major respect. In previous models, the excitation spectrum is totally specified by the fundamental frequency and a voiced/unvoiced decision for the entire spectrum. In this new model, the excitation spectrum is specified by the fundamental frequency and a voiced/unvoiced decision for each group of harmonics of the fundamental.

The excitation spectrum $E_w(\omega)$ is obtained from the fundamental frequency and the voiced/unvoiced information by combining segments of a periodic spectrum $P_w(\omega)$ in the frequency regions declared voiced with segments of a random noise spectrum in the frequency regions declared unvoiced. The periodic spectrum $P_w(\omega)$ is completely determined by the fundamental frequency. The V/UV information allows us to mix the harmonic spectrum with a random noise spectrum in a frequency-dependent manner. This model is motivated by the observation that spectra in mixed voicing segments of clean speech or in voiced segments of noisy speech tend to have regions of the spectrum dominated by harmonics of the fundamental and other regions dominated by noise-like energy. We hypothesize that humans can discriminate between frequency regions dominated by harmonics of the fundamental and those dominated by noise-like energy and employ this information in the process of separating voiced speech from random noise. Elimination of this acoustic cue in vocoders based on simple excitation models may help to explain the significant intelligibility decrease observed with these systems in noise [4].

As previously stated the new speech model assigns each group of harmonics to be either voiced or unvoiced. In [1,3] a voiced/unvoiced decision was made for each individual harmonic. However, in [2] the voiced/unvoiced information was reduced to a single decision for each group of three harmonics. This change was found to preserve the high quality capability of the MBE speech model, while substantially reducing the number of bits required to represent the voiced/unvoiced information. Using this approach noisy regions of the excitation spectrum are represented using one bit for each group of three harmonics. This is a distinct advantage over simple harmonic models [5], where noisy regions are synthesized from the coded phase requiring several bits per harmonic.

3 Speech Analysis and Synthesis

The parameters of the MBE speech model consist of the fundamental frequency, the V/UV information, and the spectral envelope. Our approach to estimating these parameters is similar to the one presented by Griffin in [2]. This approach attempts to estimate the excitation and system parameters which minimize the difference between the original and synthetic speech spectra. In general the error between the original and synthetic speech spectra can be expressed as:

$$E = \frac{1}{2\pi} \int_{-\pi}^{\pi} G(\omega) |S_w(\omega) - \hat{S}_w(\omega)|^2 d\omega \quad (1)$$

where $\hat{S}_w(\omega)$ is the synthetic speech spectrum and $G(\omega)$ is a frequency dependent weighting function. In order to find the parameter set which achieves the minimum error it is necessary to solve a highly non-linear optimization problem. For this reason we use a different approach in which we first

minimize over the spectral envelope and the fundamental frequency assuming that the speech is voiced, and then we determine the V/UV information.

The resulting method can be viewed as an analysis-by-synthesis system. For a given fundamental frequency the spectral envelope can be represented by a set of complex harmonic coefficients, which correspond to the value of the spectral envelope at the harmonics of the fundamental frequency. The harmonic coefficients which minimize the error for a given fundamental frequency are found through the solution of a set of uncoupled linear equations. This combination of fundamental frequency and harmonic coefficients can then be used to generate a synthetic spectrum which is used to evaluate (1). The resulting error is the minimum attainable for that particular fundamental frequency. By calculating this error function versus all fundamental frequencies of interest, a global minimum can be found. The V/UV decisions are made based upon the spectrum of the minimum error. We first obtain the error spectrum which is the difference between $S_w(\omega)$ and the synthetic spectrum with the minimum error. The average value of the magnitude error spectrum is then found over the region corresponding to each group of three harmonics. If this average exceeds a fixed threshold then the region is declared unvoiced, otherwise the region is declared voiced.

To synthesize speech from the estimated model parameters we use separate techniques for the voiced and unvoiced portion of the speech signal. The voiced speech is synthesized using a bank of tuned oscillators. For a particular speech segment, an oscillator is assigned to each harmonic which has been declared voiced. The amplitude, phase and frequency of each oscillator are varied over the length of each segment. Once the oscillator parameters have been calculated for each harmonic, the voiced portion of the speech signal is formed by summing the contribution from each harmonic oscillator.

In order to complete the synthesis procedure the unvoiced speech must be reconstructed. This is accomplished by calculating the spectrum of a windowed noise sequence and weighting the magnitude according to the estimated harmonic coefficients. The regions corresponding to voiced harmonics are zeroed out, so that they do not contribute any energy. The inverse transform of this spectrum is then taken and used with the weighted overlap-add procedure [6] to generate the unvoiced speech.

A detailed description of the analysis and synthesis algorithms mentioned above can be found in [2].

4 Parameter Quantization

The primary goal in the design of our 4.8 kbps MBE vocoder was to preserve the major benefits associated with the MBE speech model - the high quality speech synthesis capability and the robustness to background noise. In addition we were interested in maintaining several other properties which had been achieved by previous MBE vocoders. These included reasonable computation and storage requirements and a small coding delay. These features are important in the application of our system to real-time speech communication.

The MBE model parameters which are estimated for each frame include the fundamental frequency, the voiced/unvoiced

decisions, and a set of spectral magnitudes and phases. Transmission of these parameters at a 50 Hz frame rate was found to yield high quality speech. This allows us to use 96 bits for coding the model parameters for each frame. We have designed the system around a 4 KHz speech bandwidth, thus N , the number of harmonics per frame, is given by $N = 4000/f_0$, where f_0 is the fundamental frequency for that frame. In order to account for this variability in the number of harmonics the fundamental frequency is quantized first. A bit allocation strategy is then used to assign bits over the various other parameters. These are then quantized and transmitted as described below.

The first parameter which is quantized in each frame is the fundamental frequency. This parameter is quantized by the estimation algorithm to 1 Hz increments between 80 Hz and 500 Hz. Fixed length encoding of this value would require 420 levels or 9 bits. However, due to the slowly varying nature of speech the frame to frame deviation of the fundamental frequency is usually small. In order to exploit this fact the difference between the current fundamental frequency and the previous one is encoded. If this value lies in the range between -4 Hz and 3 Hz, then this value is coded using 4 bits. If it lies in the range between -8 Hz and -5 Hz or between 4 Hz and 7 Hz then 5 bits are used. If neither of these cases apply then 11 bits are used. This scheme results in an average of about 6 bits per frame.

The next parameters to be quantized are the voiced/unvoiced decisions. As previously mentioned a single V/UV decision is made for each group of three harmonics. This information is encoded by assigning a single bit per decision, up to a maximum of 12 bits. This is sufficient to represent the voiced/unvoiced information for the first 36 harmonics. If a frame consists of more than 36 harmonics, then the remaining ones are declared unvoiced by default.

Our system quantizes the phase of the voiced harmonics which lie in the range between 1 and 12. Informal listening tests have indicated that lack of phase information adds a reverberant quality to the synthesized speech. However, due to the small number of available bits, all of the harmonic phases cannot be coded. Since the effects of phase information were found to be more pronounced at low frequencies, we decided to code only the phase of the voiced harmonics which lie in the aforementioned range. The phase of the voiced harmonics outside this region are not coded and are randomly chosen using a uniform distribution. The phase of all unvoiced harmonics is not coded, since this information is not needed by the synthesis algorithm.

To quantize the phase, we first calculate a predicted phase based on the tuned oscillator description of voiced speech [2,3]. This phase predictor captures the information which is contained in previous frames about the phase of the current frame. The difference between the actual phase and the predicted phase forms a phase residual. This value has been found to have less entropy than the phase itself and therefore can be quantized more efficiently.

Previous MBE coding systems have quantized the phase residual using either uniform or non-uniform quantization. However, these techniques proved to be unsatisfactory for our 4.8 kbps system. In our method we separate the phase residuals into groups of 3. Each group is then block-quantized

to one of 64 levels. The level can be represented with 6 bits, yielding an average of 2 bits per phase. The results obtained from this procedure are perceptually indistinguishable from 4 bit uniform quantization and 3 bit non-uniform quantization.

The quantization of the harmonic magnitudes in our 4.8 kbps system was accomplished in a manner fundamentally different from previous MBE speech coders [2,3]. In these previous systems, the harmonic magnitudes were quantized using techniques similar to those employed in channel vocoders. Specifically, the logarithm of the magnitudes are differentially quantized across frequency. Although this technique proved satisfactory in earlier systems, it was found to introduce noticeable distortions when used in a 4.8 kbps system. This technique utilizes only a limited amount of the redundancy which exists among the harmonic magnitudes. In order to take greater advantage of this redundancy a more efficient approach was developed. This scheme is based on a new time-frequency framework which is shown in Figure 1. Every 20 ms., the harmonic magnitudes for a new speech frame are estimated. These parameters can therefore be viewed as an image, with time on one axis and frequency on the other. The frequency index corresponds to the harmonic number of the magnitude, while the time index corresponds to the frame number. Experimental results demonstrate that substantial interdependencies exist along both the time and frequency directions. In order to exploit this feature we have adopted a transform coding approach, similar to that found in image coding. The harmonic magnitudes are first sub-divided into time-frequency blocks. Each block is then transformed and the result is quantized. This technique removes substantially more redundancy than the channel coding approach. As a result fewer bits are needed for a given quality level.

The time-frequency framework which is described above can be used with a number of quantization schemes. Various transform coding and vector quantization approaches can be applied to coding each time-frequency sub-block. In addition the block size can be varied over a wide range of values. The particular choices which are made determine the system's performance in terms of quality, computation, storage and delay.

In the system which we have developed, we use sub-blocks of length 8 in the frequency direction and of length 1 in the

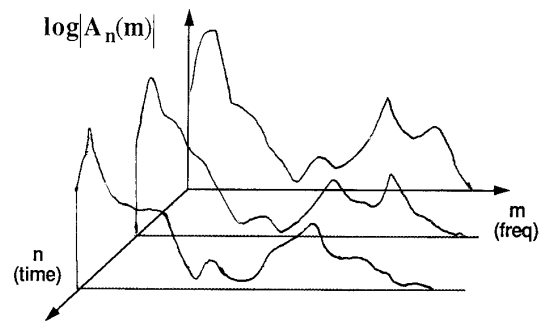


Figure 1: Time-Frequency Representation of Spectral Magnitudes

time direction. The logarithms of the harmonic magnitudes are transformed with a Discrete Cosine Transform (DCT), and the DCT coefficients are then passed through uniform quantizers. The bit allocation strategy and quantizer characteristics are based on the long term characteristics of the DCT coefficients. The total number of bits used to code the harmonic magnitudes is constrained to equal 96 minus the number of bits used to quantize the fundamental frequency, the V/UV information and the harmonic phases.

The selection of the DCT with an 8 by 1 block size was made for several reasons. The DCT is known to have good decorrelation properties. In addition an 8 by 1 DCT can be computed with a fast algorithm, yielding advantages in speed and storage. Our particular choice of the block size is also motivated by the desire to limit the coding delay, which is directly related to the time length of the block. Rather than using a two-dimensional DCT with its additional coding delay, we chose to implement a hybrid coding scheme which uses an 8 by 1 DCT along with differential quantization in the time direction. Figure 2 shows a block diagram of this hybrid approach.

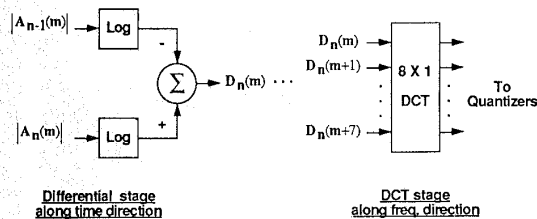


Figure 2: Hybrid Coding of Spectral Magnitudes

5 Performance Evaluation

Our 4.8 kbps speech coding system has been implemented on a SUN 3 computer in the C programming language. The system operates with a coding delay of 90 ms. In addition our current implementation runs at a rate of approximately 120 times real time for the entire analysis, coding, decoding, and synthesis process. Our new coding technique has not affected the delay, but it has increased the computational requirements by approximately 10 percent, relative to the previous 8.0 kbps system [2]. A real-time implementation of this system should be possible through the use of a special purpose DSP architecture.

Our 4.8 kbps system has been used to code a variety of sentences including both clean and noisy speech. Using these sentences, we have evaluated our system through informal listening tests. Results indicate that the quality of the coded speech is high for both clean and noisy speech. This is in sharp contrast to many other vocoder systems which develop a distinct "buzziness" in the presence of background noise. We have also compared our system with the aforementioned 8.0 kbps system and have obtained favorable results. Specif-

ically, we have found that the performance of our system is equivalent to that of the higher bit rate system in almost all cases. The primary degradation which remains is in the form of slight a reverberance due to the lack of enough coded phase information.

6 Conclusions

In this paper we have presented the development of a 4.8 kbps Multi-Band Excitation speech coder. This system was developed using several new approaches to quantize the MBE model parameters. These techniques were designed to utilize additional redundancy amongst these parameters, thereby permitting more efficient quantization. The results of informal listening tests indicate that this system can achieve high quality for both clean and noisy speech. In addition to informal listening, we are currently performing a formal Diagnostic Rhyme Test (DRT) to evaluate the performance of our 4.8 kbps system. The results of this test will be reported in our future publications.

References

- [1] Daniel W. Griffin and Jae S. Lim, "A New Model-Based Speech Analysis/Synthesis System," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 513-516, Tampa, Florida, March 26-29, 1985.
- [2] Daniel W. Griffin, "Multi-Band Excitation Vocoder," *Ph.D. Thesis*, E.E.C.S. Department, M.I.T., 1987.
- [3] Daniel W. Griffin and Jae S. Lim, "A High Quality 9.6 kbps Speech Coding System," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 125-128, Tokyo, Japan, April 13-20, 1986.
- [4] B. Gold and J. Tierney, "Vocoder Analysis Based on Properties of the Human Auditory System," M.I.T. Lincoln Laboratory Technical Report, TR-670, December 1983.
- [5] R. J. McAulay and T. F. Quatieri, "Mid-Rate Coding Based on a Sinusoidal Representation of Speech," *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, pp. 945-948, Tampa, Florida, March 26-29, 1985.
- [6] Daniel W. Griffin and Jae S. Lim, "Signal Estimation From Modified Short-Time Fourier Transform," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-32, no. 2, pp. 236-243, April 1984.